

An Evolutionary Spectrum Approach to Incorporate Large-scale Geographical Descriptors on Global Processes

Stefano Castruccio¹ and Joseph Guinness²

February 25, 2016

Abstract

We introduce a nonstationary spatio-temporal model for gridded data on the sphere. The model specifies a computationally convenient covariance structure that depends on heterogeneous geography. Widely used statistical models on a spherical domain are nonstationary for different latitudes, but stationary at the same latitude (*axial symmetry*). This assumption has been acknowledged to be too restrictive for quantities such as surface temperature, whose statistical behavior is influenced by large scale geographical descriptors such as land and ocean. We propose an evolutionary spectrum approach that is able to account for different regimes across the Earth's geography, and results in a more general and flexible class of models that vastly outperforms axially symmetric models and captures longitudinal patterns that would otherwise be assumed constant. The model can be estimated with a multi-step conditional likelihood approximation that preserves the nonstationary features while allowing for easily distributed computations: we show how the model can be fit to more than 20 million data points in less than one day on a state-of-the-art workstation. The resulting estimates from the statistical model can be regarded as a synthetic description (i.e. a compression) of the space-time characteristics of an entire initial condition ensemble.

Key words: land ocean nonstationarity, global space-time model, axial symmetry, evolutionary spectrum, climate output compression

Short title: land/ocean nonstationarity

¹School of Mathematics & Statistics, Newcastle University, Newcastle Upon Tyne, NE1 7RU United Kingdom. E-mail: stefano.castruccio@ncl.ac.uk

²Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695, United States. E-mail: joeguinness@ncsu.edu

1 Introduction

Providing efficient and flexible models for data on a spherical domain is a topic of great importance in climate science, as the statistical model can be used to fit global data. In particular, in the context of Earth System Models (ESMs), this could lead to efficient methods for compressing large quantities of data. Isotropic models have been widely acknowledged as being inadequate for data on a spherical domain (Gneiting, 2013a), and defining valid nonstationary processes is listed among the sixteen open problems in modeling spherical data in Gneiting (2013b). By regarding a random field as solution of a stochastic partial differential equation (Lindgren et al., 2011) on a spherical domain, Bolin and Lindgren (2011) proposed a nested stochastic partial differential equations approach, which yielded a field with Matérn-like covariance structure but could also be extended beyond axial symmetry to nonstationary models by allowing flexible differential operators in the stochastic equation. Jun and Stein (2007, 2008); Jun et al. (2008); Jun (2011) restrict three-dimensional isotropic fields to a sphere and apply partial derivatives with respect to latitude and longitude, obtaining a model which assumes stationarity if the data are at the same latitude, and nonstationarity otherwise (*axially symmetric* (Jones, 1963), see theoretical details in Hitczenko and Stein (2012); Huang et al. (2012)). Such models are conceptually attractive for data such as surface temperature, whose statistical properties clearly depend on latitude. Castruccio and Stein (2013) and Castruccio and Genton (2014) proposed a spectral approach that is flexible and computationally efficient when the data are on a regular grid over the sphere. This method proposes to separately consider the process by latitudinal bands, fit a Matérn-like covariance across longitudes, and then estimate the multi-band dependence, thus reducing the likelihood evaluation with the full dataset to a low dimensional parameter space. The main limitation of these models, as acknowledged in the aforementioned literature, is the assumption of stationarity in longitude at each latitude.

For physical quantities such as surface temperature, it is expected that large scale geographical descriptors such as land/ocean will impact the statistical behavior of the data. Recently Jun (2014) proposed a modified Matérn process with smoothness changing over land and ocean, which showed dramatic improvements over the axially symmetric model. The model parameters, however, were not simple to interpret given their definition through a differential operator over an isotropic process and the fitting procedure was not computationally feasible for analyzing millions of data points.

This work introduces a new class of covariance functions on spheres that includes axially symmetric models as special cases and is capable of incorporating geographic covariates into the model. For the surface temperature data we consider, the most prominent and influential large scale geographic descriptor is land versus ocean, so we focus our work on this covariate, but as we describe in Section 3.2, the ideas generalize to other covariates as well. We also propose a reformulation of the latitudinal dependence of the model in terms of a stationary AR(1) process, and introduce a nonstationary generalization which is more flexible in capturing different behaviors in the tropics.

For inference, we devise a step-wise conditional likelihood approach that fully exploits the gridded geometry of climate model output and is able to achieve a fit of more than 20 million data points in less than one day by allowing code parallelization on a state-of-the-art workstation. The proposed method vastly outperforms the axially symmetric model in terms of standard model selection metrics, and is also able to capture patterns in the longitudinal contrasts that would be otherwise assumed constant. The set of estimated parameters can then be used to almost instantaneously produce new surrogate simulations on a common laptop, thus allowing an end user to conveniently test initial scientific hypotheses on a high (spatial) resolution ensemble without downloading it, or remotely aggregating data in space or time and losing valuable information at fine scale. Besides, the estimated parameters

can be regarded as descriptors of the information for every member of the given initial ensemble (Castruccio and Genton, 2016), and thus as a compression algorithm (Rissanen, 1989; Hansen and Yu, 2001). The proposed statistical model achieves a compression rate of approximately 3:100, which is vastly superior to traditional bit-by-bit compression algorithms that can achieve at most a 1:5 ratio.

The model can also be viewed as an emulator of an initial condition ensemble (Castruccio and Genton, 2016), under the assumption that runs are independent for different initial conditions (Lorenz, 1963; Collins and Allen, 2002; Collins, 2002; Branstator and Teng, 2010). The use of emulators as data compressors is, to our knowledge, new to the climate community as they are traditionally used for calibration and sensitivity analysis (Sansó et al., 2008; Sansó and Forest, 2009; Bhat et al., 2012; Drignei et al., 2008; Chang et al., 2015) or scenario extrapolation (Holden and Edwards, 2010; Holden et al., 2013; Castruccio et al., 2014). Having statistical models (emulators) that accurately describe the model output allows us to avoid storing the entire initial condition ensemble, whose individual member requires significant storage space. This proposed methodology has shown promising results and can be generalized to multiple variables (not necessarily in the atmospheric part of the model), climate models and scenarios, and to finer temporal scales. In all these cases, the benefits of a statistical-based data compression will be even more evident as the size of the data, and consequently the expected time of downloading the full climate run, will significantly increase.

The remainder of the paper is organized as follows. Section 2 introduces the data set and discusses nonstationarity across longitude. Section 3 describes the statistical model, discusses the computational challenges that arise when fitting this with very large data sets and suggests a stepwise model-fitting approach to address these challenges in a way that exploits the geometry of the sphere. Section 4 shows the comparison with the axially

symmetric model. Section 5 shows how the fitted model can be used to compress the initial condition ensemble and how to generate surrogate runs from the estimated parameters. Section 6 concludes with a discussion.

2 The CMIP5-CCSM4 ensemble

The Coupled Model Intercomparison Project phase 5 (CMIP5 Taylor et al., 2012) is a set of coordinated experiments from many modeling groups to provide uniform and comparable assessment of climate response under different climate models for the latest IPCC Assessment Report (IPCC, 2013). In this work we focus on the NCAR Community Climate System Model 4 (CCSM4 Gent et al., 2011), under the Representative Concentration Pathway 8.5 (rcp85 Van Vuuren et al., 2011) from 2006 to 2100, for a total of 95 years. We consider annual temperature at surface (considered at a standard height of 2 meters above ground level), which is on a regular 192×288 grid over latitude and longitude. We remove the bands near the poles (south of 62 degrees south and north of 70 degrees north) so that each spatial field consists of 142×288 points. The removal of the Arctic and Antarctic bands was performed to avoid inference on latitudes where two locations at one longitudinal lag were very close, consistently with Castruccio and Stein (2013); Castruccio and Genton (2014) and Castruccio and Genton (2016). This would have led to very smooth spatial processes, and consequently computational challenges that would have added to the already substantial complexity of the inference scheme. Under rcp85, the CCSM4 was run under 6 different sets of initial conditions, therefore generating 6 independent realizations (Lorenz, 1963; Collins and Allen, 2002; Collins, 2002; Branstator and Teng, 2010). The total size of the dataset is therefore $142 \times 288 \times 95 \times 6 = 23.3$ million points. The movie `movie_cm.avi` in the supplementary material shows a realization of this climate model. The annual temperature shows evidence of normality, as reported in the supplementary material.

We denote by \mathbf{T}_r the temperature for realization $r = 1, \dots, R$, by $L_m \in (-\pi/2, \pi/2)$, $m = 1, \dots, M$ the latitude, by $\ell_n = 2\pi n/N$, $n = 1, \dots, N$ the longitude, by t_k , $k = 1, \dots, K$ the year, where $R = 6$, $M = 142$, $N = 288$ and $K = 95$. Thus, the temperature for realization r is represented as

$$\mathbf{T}_r = \{\mathbf{T}_r(L_1, \ell_1, t_1), \dots, \mathbf{T}_r(L_M, \ell_1, t_1), \mathbf{T}_r(L_1, \ell_2, t_1), \dots, \mathbf{T}_r(L_M, \ell_N, t_K)\}.$$

The defining assumption of axially symmetric models is that the process is stationary across longitude at each latitude. In this work, we relax this assumption to allow geographic covariates to be incorporated into the covariance function and inform more complex spatial dependence structures. Local geography can have a strong impact on the statistical characteristics of surface temperature data, so a natural deviation from the stationary assumption is to allow the statistical properties of the process to differ over land and ocean, which is the most dramatic geographic descriptor at large scales. A simple modeling solution is to divide the temperature at each location (L_m, ℓ_n) by the standard deviation s_{L_m, ℓ_n} obtained from a simulation from the same climate model but with no forcing (a *control run* in geoscience terminology), as proposed in Castruccio and Stein (2013), to obtain more realistic conditional simulations. While producing improved results, this approach does not allow for a changing correlation structure across longitude, and in particular across land and ocean. Indeed, empirical estimates of the second-order (covariance) structure show a strong dependence on the land/ocean variable. To see this, we consider the difference between two realizations $\mathbf{T}_1 - \mathbf{T}_2$ (to remove any trend), normalize it by s_{L_m, ℓ_n} , and compute

$$|\hat{f}_{L_m}^j(c)|^2 = \frac{1}{K} \sum_{k=1}^K \frac{p(h^j)}{N} \left| \sum_{n=1}^N h^j(\ell_n) \frac{\mathbf{T}_1(L_m, \ell_n, t_k) - \mathbf{T}_2(L_m, \ell_n, t_k)}{s_{L_m, \ell_n}} e^{-i\ell_n c} \right|^2, \quad (1)$$

for $j = 1, 2$, which is the periodogram of a tapered version of the data averaged over time at latitude L_m , where the taper h^1 is a smooth function that is equal to zero when $\ell_n \in \text{ocean}$, and h^2 is a smooth function that is equal to zero when $\ell_n \in \text{land}$, and $p(h^j)$ is a normalizing

constant. Thus we can view $|\hat{f}_{L_m}^1|^2$ as a periodogram for the land data averaged over time at latitude L_m , and $|\hat{f}_{L_m}^2|^2$ as a periodogram for the ocean data averaged over time at latitude L_m . In Figure 1, we plot $\log(|\hat{f}_{L_m}^j|^2)$ at latitude $L_m = 41^\circ$. Because the two log periodograms in Figure 1 are not parallel—which would indicate similar correlation structure over land and ocean—it is clear that the data exhibit land/ocean nonstationary correlation, and that the process over the ocean is much smoother than the process over the land at this latitude.

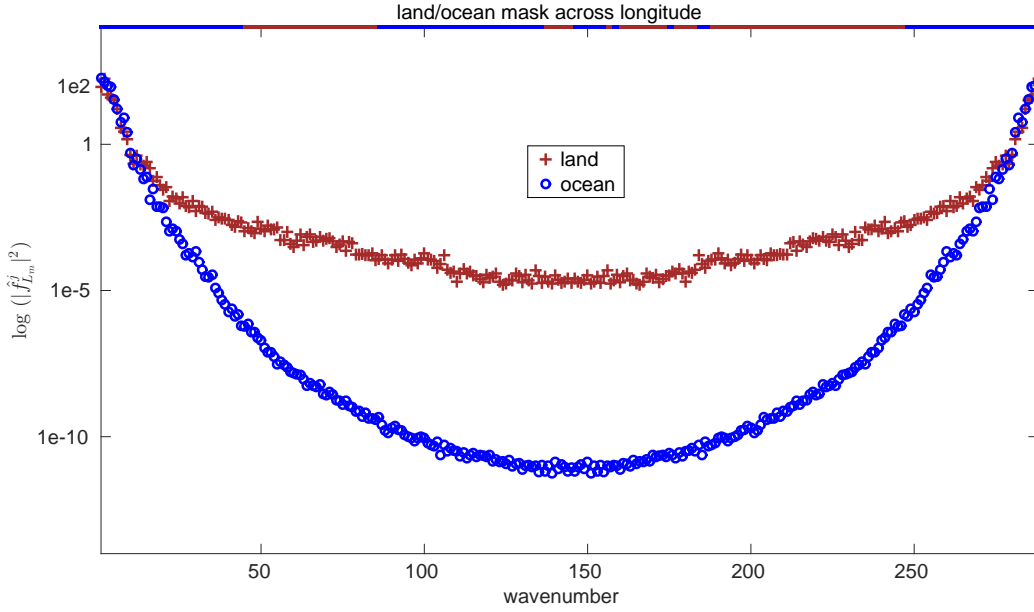


Figure 1: Comparison of the land/ocean periodogram of the difference between two realizations of rcp85 computed with (1), each pixel being normalized by its standard deviation from the control run. The latitude band represented is $\approx 41^\circ\text{N}$, and the periodogram is averaged over all years. On the top the land/ocean mask across longitude is represented.

The land and ocean periodograms in Figure 1 motivate the use of a model that allows for a different behavior across land and ocean for the global space-time temperature data. In this work, we define a nonstationary model with changing spectrum across these two domains, denoted as *evolutionary spectrum*. The details are provided in Section 3.2

3 The space time model

In this section we describe the global space-time model. We first introduce in Section 3.1 a fundamental result that allows us to fit the stochastic component of the statistical model without defining a parametric form for the mean. In Section 3.2 we describe the model. Section 3.3 shows the results and discusses the computational challenges of fitting the proposed model on a dataset with tens of millions of data points.

3.1 Preliminaries

Denote by $\mathbb{E}(\mathbf{T}_r) = \boldsymbol{\mu}$ the mean temperature across realizations. Since realizations differ just in their initial condition, and since climate models tend to forget their initial state after a short number of temporal steps (Lorenz, 1963; Collins and Allen, 2002; Collins, 2002; Branstator and Teng, 2010), we can assume that the space-time field \mathbf{T}_r is independent across r :

$$\mathbf{T}_r = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_r, \quad \boldsymbol{\varepsilon}_r \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \quad (2)$$

where $\boldsymbol{\theta}$ is a vector of unknown covariance parameters. The noticeable advantage of having independent realizations is that $\boldsymbol{\theta}$ can be estimated without any parametrization of $\boldsymbol{\mu}$ via restricted loglikelihood. Castruccio and Stein (2013) proved the following result, which formulates the restricted loglikelihood for \mathbf{T}_r in a computationally convenient form.

Result 1 *Denote with $\frac{1}{R} \sum_{r=1}^R \mathbf{T}_r = \bar{\mathbf{T}}$ the average temperature across realizations. Let $\mathbf{D}_r = \mathbf{T}_r - \bar{\mathbf{T}}$. The negative restricted loglikelihood for (2) is*

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{D}) &= \frac{KNM(R-1)}{2} \log(2\pi) + \frac{1}{2}(R-1) \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| \\ &\quad + \frac{1}{2}KNM \log(R) - \frac{1}{2} \sum_{r=1}^R \mathbf{D}_r^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{D}_r. \end{aligned} \quad (3)$$

Also, the corresponding estimator for $\boldsymbol{\mu}$ obtained by generalized least squares is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{T}}$.

Throughout this work we make use of (3) to estimate the space/time structure of the data.

3.2 Sphere-Time Covariance

Denote by $\boldsymbol{\varepsilon}(t_k; r) = \{\boldsymbol{\varepsilon}_r(L_1, \ell_1, t_k), \dots, \boldsymbol{\varepsilon}_r(L_N, \ell_M, t_k)\}$ the vector of the stochastic term of (2) at time t_k , by $\mathbf{T}(t_k; r)$ the temperature at year t_k for realization r and by $\mathbf{D}(t_k; r) = \mathbf{T}(t_k; r) - \bar{\mathbf{T}}(t_k; r)$. We assume that $\boldsymbol{\varepsilon}(t_k; r)$ is correlated across time, and previous work (Castruccio and Genton, 2014) has shown that an AR(2) model with different coefficients for every grid point is sufficiently flexible, as no evidence of cross-temporal dependence or nonseparability between space and time was found on annual scale:

$$\boldsymbol{\varepsilon}(t_k; r) = \boldsymbol{\Phi}_1 \boldsymbol{\varepsilon}(t_k - 1; r) + \boldsymbol{\Phi}_2 \boldsymbol{\varepsilon}(t_k - 2; r) + \mathbf{S} \mathbf{H}(t_k; r), \quad (4)$$

where $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ are two $NM \times NM$ diagonal matrices with the autoregressive coefficients for each location, and \mathbf{S} is a $NM \times NM$ diagonal matrix with the associated standard deviations.

The unscaled innovations $\mathbf{H}_r(L_m, \ell_n, t_k)$ are independent across time and describe the spatial dependence across the sphere. We propose to model the process in the spectral domain across longitudes, and then to model the dependence across latitudes:

$$\begin{aligned} \mathbf{H}_r(L_m, \ell_n, t_k) &= \sum_{c=0}^{N-1} f_{L_m, \ell_n}(c) e^{i\ell_n c} \tilde{\mathbf{H}}_r(c, L_m, t_k), \\ \text{corr} \left\{ \tilde{\mathbf{H}}_r(c, L_m, t_k), \tilde{\mathbf{H}}_{r'}(c', L_{m'}, t_{k'}) \right\} &= \mathbf{1}\{c = c', k = k', r = r'\} \rho_{L_m, L_{m'}}(c), \end{aligned} \quad (5)$$

where c is a wavenumber. If $f_{L_m, \ell_n} = f_{L_m}$ for every L_m , then this would be a standard stationary model with spectral density $|f_{L_m}|^2$. This proposed model assumes that the spectral density is not exactly constant in longitude, but evolves (hence the term *evolutionary spectrum*) according to $|f_{L_m, \ell_n}|^2$. Models with evolutionary spectra allow us to flexibly specify the local covariance properties at every location while ensuring that the resulting covariance function is positive definite. Evolutionary spectra were first introduced by Priestley (1965) to model nonstationary time series data, and Guinness and Stein (2013) provided computationally efficient methods for fitting models with evolutionary spectra to nonstationary time

series data. In this work, we adapt the evolutionary spectra to model nonstationarity over longitude rather than over time, which requires a discrete spectrum because of the circular domain on which the data at a single latitude fall.

Certain regularity conditions can be imposed on f_{L_m, ℓ_n} near the poles to achieve mean square continuity (see supplementary material). $\rho_{L_m, L_{m'}}$ is the correlation in the spectral domain (or *coherence* in spectral analysis) between latitudes L_m and $L_{m'}$ among $\tilde{\mathbf{H}}_r$ for the same wavenumber, time and realization. Alternatively, one could deviate from the stationary assumption across longitude by introducing dependence across wavenumber in $\tilde{\mathbf{H}}_r$. Our choice to use evolutionary spectra to model the nonstationarities is motivated by the need to include geographic covariates.

The typical approach (Priestley, 1965) for $f_{L_m, \ell_n}(c)$ is to describe the dependence on ℓ_n according to covariates $X^j(L_m, \ell_n)$ as

$$f_{L_m, \ell_n}(c) = \sum_{j=1}^p f_{L_m}^j(c) X^j(L_m, \ell_n). \quad (6)$$

In this work, we propose a novel model where land and ocean are included as covariates to allow for different statistical behaviors across these two domains. Thus, $f_{L_m, \ell_n}(c)$ can be expressed as

$$f_{L_m, \ell_n}(c) = f_{L_m}^1(c) b_{\text{land}}(L_m, \ell_n) + f_{L_m}^2(c) \{1 - b_{\text{land}}(L_m, \ell_n)\}, \quad (7)$$

where the component spectra are modeled according to the parametric form

$$|f_{L_m}^j(c)|^2 = \frac{\phi_{L_m}^j}{\{(\alpha_{L_m}^j)^2 + 4 \sin^2\left(\frac{c}{N}\pi\right)\}^{\nu_{L_m}^j + 1/2}}, \quad j = 1, 2, \quad (8)$$

and b_{land} is a function between 0 and 1 that modulates the relative contribution of the land regime. (8) is a Matérn-like spectrum, which is modified for the case of data on a circle to allow for a smooth transition at high wavenumbers and has been shown to adequately capture the longitudinal behavior of temperature at surface for different latitudes better

than the traditional Matérn-like spectrum (Castruccio and Stein, 2013; Poppick and Stein, 2014).

Choosing an indicator function for b_{land} would result in abrupt transitions between the two regimes at land/ocean boundaries and in misfit of the data, as shown in the supplementary material. We therefore introduce a smoother taper function to transition between land and ocean:

- Let $I_m(\ell_n)$ denote the indicator function of land at latitude L_m and longitude ℓ_n .

Wherever there is a land/ocean transition, we modify $I_m(\ell_n)$ so that is equal to one for g_{L_m} more grid points, where g_{L_m} is an integer number that can also be negative.

The modified indicator is denoted by $\tilde{I}_m(\ell_n; g_{L_m})$.

- Compute the Tukey taper function (Tukey, 1967) with range γ_{L_m} :

$$w_m(\ell_n; \gamma_{L_m}) = \begin{cases} \frac{1}{2} \left[1 + \cos \left\{ \frac{2\pi}{\gamma_{L_m}} (\ell_n - \gamma_{L_m}/2) \right\} \right], & 0 \leq \gamma_{L_m} < \frac{\gamma_{L_m}}{2}, \\ 1, & \gamma_{L_m}/2 \leq \ell_n < 1 - \gamma_{L_m}/2, \\ \frac{1}{2} \left[1 + \cos \left\{ \frac{2\pi}{\gamma_{L_m}} (\ell_n - 1 - \gamma_{L_m}/2) \right\} \right], & 1 - \gamma_{L_m}/2 \leq \ell_n \leq 2\pi. \end{cases} \quad (9)$$

- Convolve $\tilde{I}_m(\ell_n; g_{L_m})$ with $w_m(\ell_n; \gamma_{L_m})$:

$$b_{\text{land}}(L_m, \ell_n; g_{L_m}, \gamma_{L_m}) = \sum_{n'=1}^N \tilde{I}_m(\ell_n; g_{L_m}) w_m(\ell_n - \ell_{n'}; \gamma_{L_m}). \quad (10)$$

This formulation imposes a symmetric land/ocean transition (i.e. land/ocean and ocean/land transitions are equally smooth); however, more sophisticated models with asymmetric transitions have been tested but have not yielded significantly better results. Similarly, no significant improvements have been observed if a different taper is used (as shown in the supplementary material) or g_{L_m} and γ_{L_m} are assumed different across oceans. If we constrain

$$\phi_{L_m}^1 = \phi_{L_m}^2, \alpha_{L_m}^1 = \alpha_{L_m}^2, \nu_{L_m}^1 = \nu_{L_m}^2, \quad (11)$$

then in (5) $f_{L_m, \ell_n} = f_{L_m}$ and the model becomes stationary across longitude.

Castruccio and Stein (2013); Castruccio and Genton (2014, 2016) propose the following parametric model for $\rho_{L_m, L_{m'}}(c)$ in (5):

$$\rho_{L_m, L_{m'}}(c) = \rho_{L_m - L_{m'}}(c) = \left[\frac{\xi}{\{1 + 4 \sin^2(\frac{c}{N}\pi)\}^\tau} \right]^{|L_m - L_{m'}|} = \varphi(c)^{|L_m - L_{m'}|}. \quad (12)$$

with $\varphi(c) = \frac{\xi}{\{1 + 4 \sin^2(\frac{c}{N}\pi)\}^\tau}$. This process is equivalent to the following AR(1) process in latitude:

$$\begin{aligned} \tilde{\mathbf{H}}_{L_m}(c) &= \begin{cases} \varphi(c)\tilde{\mathbf{H}}_{L_{m-1}}(c) + \mathbf{e}_{L_m}(c), & m = 2, \dots, M, \\ \mathbf{e}_{L_1}(c) \sim \mathcal{N}(0, 1), & m = 1, \end{cases} \\ \mathbf{e}_{L_m} &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1 - \varphi(c)^2), \quad m > 1, \end{aligned} \quad (13)$$

where $\text{var}(\mathbf{e}_{L_m}(c)) = 1 - \varphi(c)^2$ to allow unit variance on $\tilde{\mathbf{H}}_{L_m}(c)$. While the coherence in (12) has been previously used in literature, the formulation of the latitudinal dependence in terms of an autoregressive process has never been acknowledged.

The formulation of the dependence as a stationary AR(1) process allows for generalization to nonstationary latitudinal processes to increase the model flexibility. In particular, in addition (12) we also propose a novel nonstationary AR(1) model for the coherences, with latitudinally-varying autoregressive parameters, that is

$$\varphi_{L_m}(c) = \frac{\xi_{L_m}}{\{1 + 4 \sin^2(\frac{c}{N}\pi)\}^{\tau_{L_m}}}. \quad (14)$$

Our diagnostics have shown that the coherences are nearly stationary outside of the tropics, so we fit nonstationary coherences within $-23^\circ < L < 23^\circ$ (i.e. in the tropics), while we assume a constant outside this region.

Thus, the model consists of three sets of parameters to be estimated

- The temporal parameters, consisting of all the entries in Φ_1 , Φ_2 and \mathbf{S} in (4), which will be denoted as θ_{time} .
- The longitudinal parameters, consisting of $(\phi_{L_m}^j, \alpha_{L_m}^j, \nu_{L_m}^j)$ in (8) and (g_{L_m}, γ_{L_m}) in (10) for $m = 1, \dots, M$. We denote the collection of all parameters as θ_{lon} .

- The latitudinal parameters, consisting of (ξ_{L_m}, τ_{L_m}) in (14) $m = 1, \dots, M$. We denote them as $\boldsymbol{\theta}_{\text{lat}}$.

3.3 Model fit and computational considerations

Despite the computationally convenient form in (3) (which can be further simplified as shown in the supplementary material) it is not feasible to perform a global optimization, since this would imply maximizing the likelihood over more than 100,000 parameters (the temporal part requires 3 parameters for each of the $142 \times 288 \approx 41,000$ locations, the spatial part requires a total number of parameters shown in the first row of Table 1). We therefore propose successive conditional approximations of (3) by assuming independence across increasingly large subsets, each approximation assuming the parameters from previous steps to be known and fixed.

1. Estimate the temporally autoregressive parameters $\boldsymbol{\theta}_{\text{time}}$, assuming that the innovations $\mathbf{H}(t_k; r)$ are independent across latitude and longitude.
2. Consider $\boldsymbol{\theta}_{\text{time}}$ fixed at their estimated values and estimate $\boldsymbol{\theta}_{\text{lon}}$, assuming the innovations $\mathbf{H}(t_k; r)$ are independent across latitudes.
3. Consider $\boldsymbol{\theta}_{\text{time}}$ and $\boldsymbol{\theta}_{\text{lon}}$ fixed at their estimated values and estimate $\boldsymbol{\theta}_{\text{lat}}$.

The choice of the blocks in the approximation, as well as the approximation order is dictated by the geometry of the problem as well as from physical considerations. Estimating the temporal structure for each location assuming no cross-correlation allows for a very fast (and scalable) computation of approximation 1. The choice of latitudinal bands in approximation 2 allows flexible estimation of the statistical parameters across latitude, which is the main descriptor of surface temperature. Further, this choice results in exactly circulant matrices across longitudes in the axially symmetric case, a feature that allows very fast computations in the spectral domain. This conditional approximation scheme can be generalized to

allow for vertical profile of temperatures as in Castruccio and Genton (2016), and can also be applied to any large space-time data set where the geometry, as well as the physics of the problem suggest the blocks, e.g. regions of interest in functional Magnetic Resonance Imaging (Castruccio et al., 2016).

The sequential model-fitting procedure can also be used to fit axially symmetric versions of the model for the innovations. This involves imposing the constraint (11) in approximation 2. In Figure 2 we see a comparison of the evolutionary spectrum model with the axially symmetric model (i.e. with constraint (11)) in terms of estimated parameters and loglikelihood. Figure 2a-c shows how land and ocean parameter estimates for the evolutionary spectrum are very different from those of the stationary model, and how there is a consistent difference across latitude. Figure 2c shows how the smoothness parameter is smaller for land than for ocean which implies, as noticed in Figure 1, that ocean temperatures tend to have a smoother behavior across the same band compared to land. Given the very large size of the data set, the parameter estimates are very precise and the estimated standard deviations¹ are two orders of magnitude smaller than the point estimates. Thus, we chose not to report the confidence intervals as they are very small compared to the differences across latitudes. We also report in Figure 2d the individual maximum loglikelihoods for each band. The loglikelihood shows a noticeable improvement for the evolutionary spectrum approach, especially in the Southern Hemisphere. In latitudes where there is no land, such as the southernmost bands considered (we removed the Antarctic regions) the evolutionary spectrum and the axially symmetric model are the same and thus have the same loglikelihood.

Approximation 3 would require estimating ξ_{L_m} and τ_{L_m} when $-23^\circ < L < 23^\circ$ and a constant value for both parameters outside the tropics, for a total of 50 parameters. Since a likelihood maximization for such number of parameters and with tens of millions of data

¹computed at each step conditional to the previous steps. For example, the standard deviations for $(\hat{\phi}_L^j, \hat{\alpha}_L^j, \hat{\nu}_L^j)$ are estimated conditional on the temporal parameters.

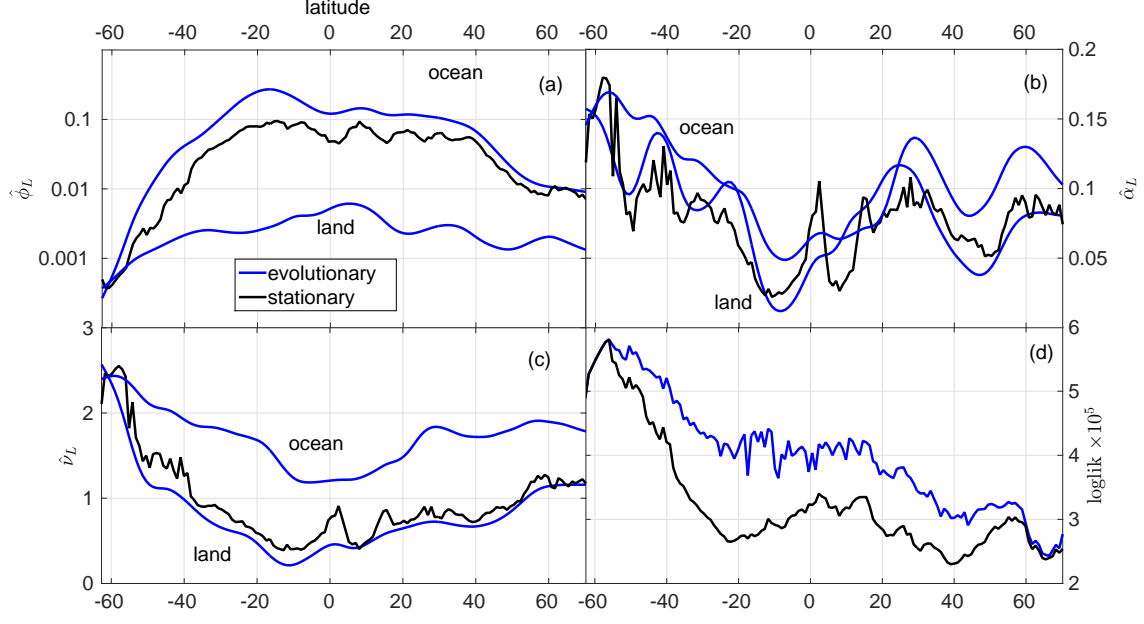


Figure 2: Comparison of the models with evolutionary spectrum (7) and the axially symmetric model with constraint (11) in terms of (a) $\log(\hat{\phi}_{L_m})$ and $\log(\hat{\phi}_{L_m}^j)$ for $j = 1, 2$, (b) $\hat{\alpha}_{L_m}$ and $\hat{\alpha}_{L_m}^j$, (c) $\hat{\nu}_{L_m}$ and $\hat{\nu}_{L_m}^j$, and (d) loglikelihood. A smoothing spline has been applied to the estimated parameters for the evolutionary spectrum approach in a-c since the pattern were less regular.

points is not feasible, we first consider (14) for adjacent bands, obtain their estimates independently for every pair of bands, which we denote as $\hat{\xi}_{L_m}^{(2)}$ and $\hat{\tau}_{L_m}^{(2)}$, and consider these estimates as fixed in approximation 3. (Since every band is involved in two fits, by convention at latitude L_m we plug in the estimates from bands (L_m, L_{m+1})). The fitted parameters for (13) and (14), along with $\hat{\xi}_{L_m}^{(2)}$ and $\hat{\tau}_{L_m}^{(2)}$, can be found in Figure S4 in the supplement. The stationary model shows some misfit, especially for ξ : this is due to model assuming a constant value across latitude for the coherence, while this is significantly smaller in the southernmost regions and at some tropical latitudes. The parameters of the nonstationary AR(1) model (14) instead are fixed and equal to $\hat{\xi}_{L_m}^{(2)}$ and $\hat{\tau}_{L_m}^{(2)}$ in the tropical regions (by construction) and, while still not capturing nontrivial latitudinal patterns outside the tropics, it results in a larger and more satisfactory estimate for ξ .

4 Model Comparison

Table 1 shows a comparison among a model that assumes spatial independence (denoted *ind*), the axially symmetric model (denoted *ax*), a model with land/ocean evolutionary spectrum with a stationary AR(1) latitudinal process (12) (denoted *ev-st*) and one with a nonstationary latitudinal AR(1) process (14) (denoted *ev-nst*).

Table 1: Comparison between different models in terms of number of parameters (excluding the temporal ones), computational time (hours), normalized restricted loglikelihood (3), and Bayesian Information Criterion (Schwarz, 1978).

Model	<i>ind</i>	<i>ax</i>	<i>ev-st</i>	<i>ev-nst</i>
# param	0	428	1138	1234
time (hours)	1.4	1.5	13.8	14.8
$\Delta\loglik/NMT(R-1)$	-2.87	-0.61	-0.0018	0
$BIC \times 10^8$	-0.1677	-1.0465	-1.2832	-1.2839

The model assuming independence is clearly the fastest to fit, as once the temporal part is estimated, the full likelihood can be evaluated just once. The axially symmetric model requires spatial parameters, but the computational time is almost equivalent and the improvement both in terms of normalized likelihood and BIC is noticeable. The evolutionary spectrum model requires approximately three times more parameters than the axially symmetric model and a noticeable increase of computational time (mostly because of the 2-band step). The resulting model, however, shows a dramatic improvement; the loglikelihood improves by 0.6 units per observation, and improves the BIC despite the large increase in the number of parameters. *ev-nst* requires more parameters (the plug-in estimates $\hat{\xi}_{L_m}^{(2)}$ and $\hat{\nu}_{L_m}^{(2)}$ at the equator) and there is small indication of a further improvement in the fit.

To assess the quality of the fit, we compute the contrast variance

$$\begin{aligned}
\Delta_{\text{ew};m,n} &= \frac{1}{KR} \sum_{k=1}^K \sum_{r=1}^R \left\{ \hat{\mathbf{H}}_r(L_m, \ell_n, t_k) - \hat{\mathbf{H}}_r(L_m, \ell_{n-1}, t_k) \right\}^2, \\
\Delta_{\text{ns};m,n} &= \frac{1}{KR} \sum_{k=1}^K \sum_{r=1}^R \left\{ \hat{\mathbf{H}}_r(L_m, \ell_n, t_k) - \hat{\mathbf{H}}_r(L_{m-1}, \ell_n, t_k) \right\}^2,
\end{aligned} \tag{15}$$

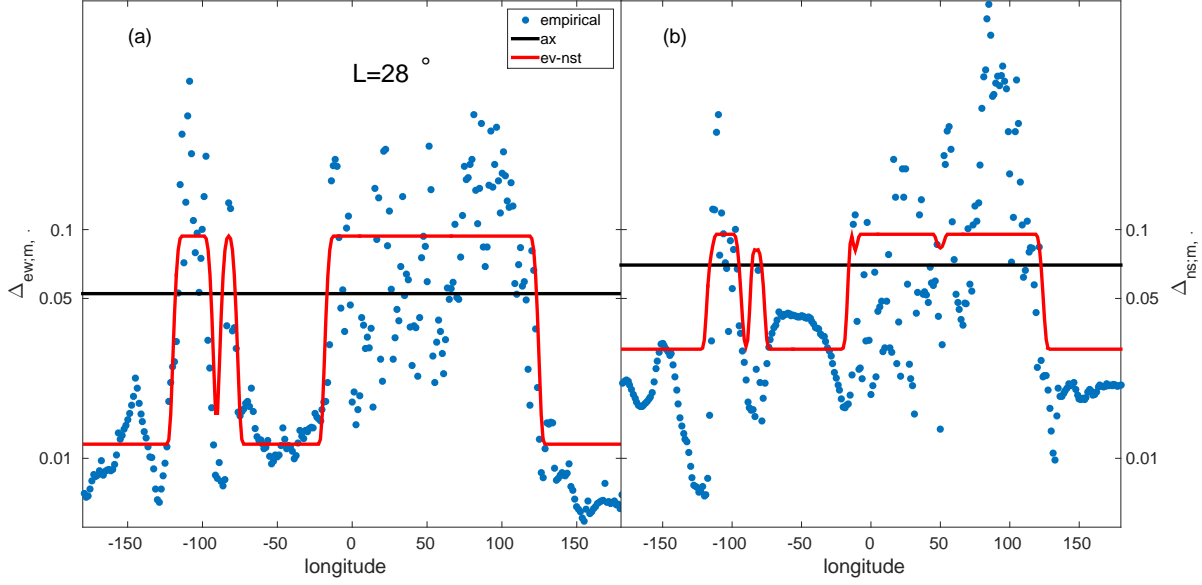


Figure 3: Estimated and fitted variances for the contiguous differences at approximately 28° North for different longitudes, averaged across time and realizations. (a): $\Delta_{ew;m,\cdot}$ and (b): $\Delta_{ns;m,\cdot}$ as in (15). The vertical axis is plotted on a log scale.

where ew=east-west and ns=north-south, and compare them with the corresponding fitted quantities according of *ax* (axially symmetric) and *ev-nst* (latitudinally nonstationary evolutionary spectrum). The result for a chosen band at approximately 28° North is shown in Figure 3. Both panels show the limits of axial symmetry which, assuming longitudinal stationarity, results in a constant value across longitude. This is clearly not adequate for temperature data, as there are significant longitudinal patterns generated by different land/ocean domains; for this latitude, $\Delta_{ew;m,\cdot}$ is nearly ten times larger over land than it is over ocean. The evolutionary spectrum model proposed here is noticeably more flexible and is able to accurately capture the changes across longitude in the contrast variances. It is apparent how different domains have different behaviors, and thus different spatial correlation, and how the fitted *ev-nst* allow for a smoother spatial behavior over the ocean. The evolutionary spectrum proposed is particularly effective in capturing $\Delta_{ew;m,\cdot}$ in Figure 3a, while some misfit is present in the Pacific Ocean for the north-south contrast variances in

Figure 3b.

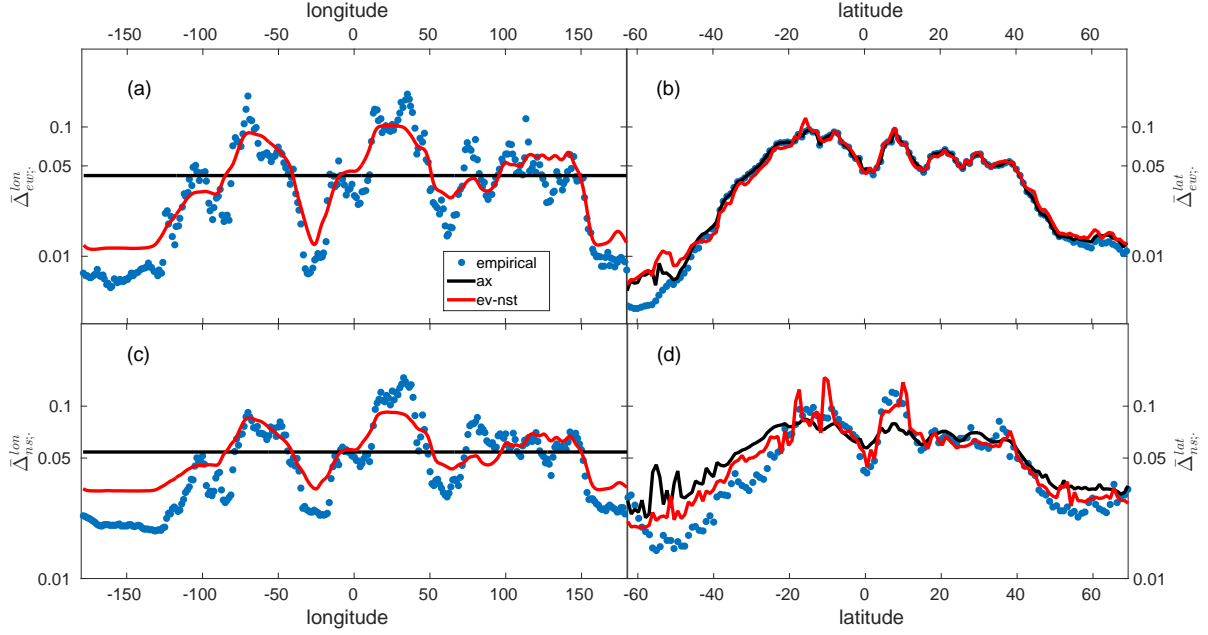


Figure 4: Estimated and fitted averaged contrast variances as in (16). (a): $\bar{\Delta}_{ew;\cdot}^{\text{lon}}$ (b): $\bar{\Delta}_{ew;\cdot}^{\text{lat}}$, (c): $\bar{\Delta}_{ns;\cdot}^{\text{lon}}$ and (d): $\bar{\Delta}_{ns;\cdot}^{\text{lat}}$. The vertical axis is plotted on a log scale.

To assess the fit for multiple latitudes, we compute the average contrast variance

$$\begin{aligned}\bar{\Delta}_{j;m}^{\text{lat}} &= \frac{1}{N} \sum_{n=1}^N \Delta_{j;m,n}, \\ \bar{\Delta}_{j;n}^{\text{lon}} &= \frac{1}{M} \sum_{m=1}^M \Delta_{j;m,n},\end{aligned}\tag{16}$$

where $j = \{\text{ew}, \text{ns}\}$. In Figure 4a-b, the values for $j = \{\text{ew}\}$ contrasts are shown, and while both *ax* and *ev-nst* are able to capture longitudinally averaged variances in panel (b) (apart from a misfit of both models in the southernmost bands), only *ev-nst* is able to capture the pattern in latitudinally averaged variance in (a), since *ax* assumes constant variance across longitudes. Figure 4c-d shows the values for $j = \{\text{ns}\}$. Similar remarks as in the previous case hold, but *ev-nst* does not fully capture the patterns of latitudinally averaged variance in panel (c), while the two models performs similarly (and with some degree of misfit in the southernmost latitudes) in longitudinally averaged variances in panel (d).

5 Simulating the initial condition ensemble

We now proceed with simulating surrogate (emulated) runs according to the evolutionary spectrum with nonstationarity in latitude (14). From (3), the mean can be estimated as $\hat{\boldsymbol{\mu}} = \bar{\mathbf{T}}$. For each location, we fit a cubic polynomial smoothing spline $\tilde{\mathbf{T}}_{m,n}$ from $\lambda \sum_{k=1}^K \{\bar{\mathbf{T}}(L_m, \ell_n, t_k) - \tilde{\mathbf{T}}_{m,n}(t_k)\}^2 + (1 - \lambda) \sum_{k=1}^K \left| \frac{d^2 \tilde{\mathbf{T}}_{m,n}}{dk^2}(t_k) \right|^2$ with mild penalty term $\lambda = 0.01$ since the climate is slowly varying (Castruccio and Genton, 2016), and we denote by $\tilde{\mathbf{T}} = (\tilde{\mathbf{T}}_{1,1}, \dots, \tilde{\mathbf{T}}_{M,N})$. To generate a simulation, the following steps are required:

- generate $\mathbf{e}_{L_m}(c) \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1 - \varphi_{L_m}(c)^2)$ with $\varphi_{L_m}(c)$ as in (14),
- compute $\tilde{\mathbf{H}}_{L_m}(c)$ with (13) and (14),
- compute $\mathbf{H}_r(L_m, \ell_n, t_k)$ with (5),
- compute $\boldsymbol{\varepsilon}_r$ with (4),
- obtain the surrogate run as $\tilde{\mathbf{T}} + \boldsymbol{\varepsilon}_r$.

Once the parameters have been estimated, a common laptop can generate hundreds of surrogate runs almost instantaneously with the aforementioned steps.

In Figure 5 we show a comparison of the six runs from the climate model ensemble and the surrogate runs in terms of temperature series near London. In panel (a), the six climate model runs are compared with six surrogate runs (offset by 2 C° to avoid superimposition). The two groups show the same trend and the same variance, but the statistical model allows to generate more runs, so that it is possible to have a better assessment of the temperature uncertainty at a given year. (In this context, by “uncertainty” we mean “uncertainty due to initial conditions”. We do not consider the uncertainty due to physical parameter calibration or forcing scenario.) In panel (b), we see how having just six climate model runs is poorly

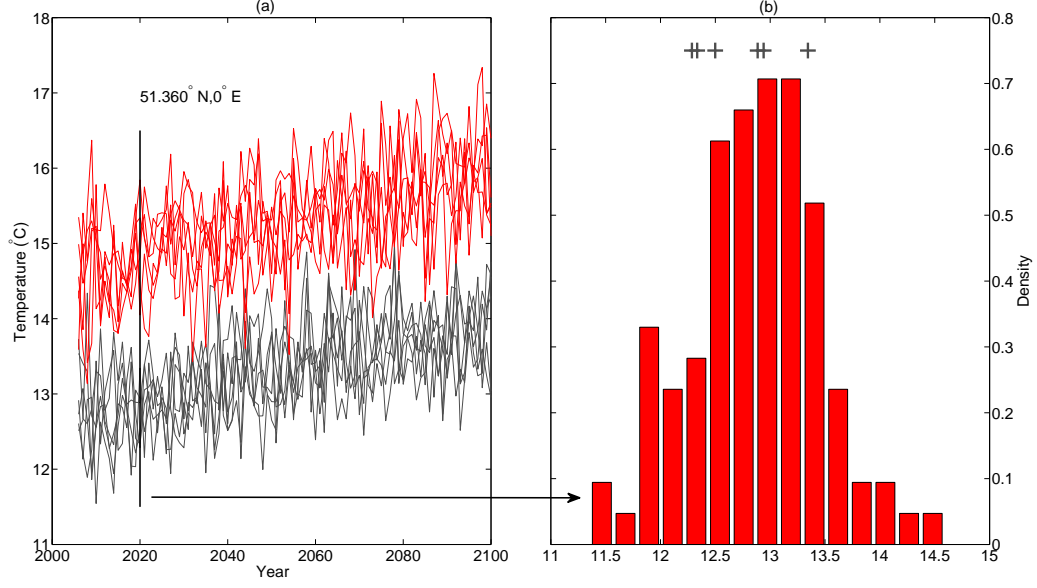


Figure 5: Comparison of climate model output with surrogate runs. (a) The six realizations of the climate model near London (in gray) are shown against six surrogate runs (in red, offset by 2 C°). (b) Histogram of the distribution of temperature for the year 2020 for the 100 surrogate runs. The gray crosses above represent the six realizations from the ensemble for the same year.

informative of the projection uncertainty for 2020, while with 100 surrogate runs it is possible to have a better assessment.

Although a comparison on a single location does not inform about the ability of the statistical model to capture the spatial variability, it is possible to produce animations of surrogate runs to detect if the spatial patterns are qualitatively consistent. Genton et al. (2015) have discussed in detail how climate model output and statistical surrogates can be compared in the case of three dimensional annual temperatures by using a virtual reality environment. In this work, we produce movies for one climate model run and a surrogate run (both in the supplementary material), which qualitatively shows similar large-scale features.

6 Conclusion and discussion

In this work we introduced a new class of spectral models that is able to incorporate geographical information to capture the nonstationary behavior of global data across longitude. We further introduced a nonstationary structure across latitude that allows for a more flexible and general description of the dependence among different bands. The evolutionary spectrum model we developed vastly outperforms axially symmetric models, showing improved performance under common model selection metrics and the estimation of the contrast variances. By using appropriate diagnostics, we show how this model is able to capture patterns across longitude that would be constant under the assumption of axial symmetry. The proposed model can be also used to incorporate further geographical information, such as orography, or can be applied to other physical quantities whose dynamics are known to be influenced by large scale geographical features, such as precipitation or winds.

The likelihood of the proposed model can be written in a computationally convenient form, which is almost as fast as in the axially symmetric case and can be successively approximated with a highly parallelizable algorithm while still preserving the main space-time structure, as shown in the diagnostics. While in this work the approximation blocks and the order of approximation (time, longitude, latitude) have been suggested by the particular problem, the multi-step approximation presented can be applied to any large space-time data set where the nature of the problem suggests blocks: for example, in functional Magnetic Resonance Imaging the brain can be naturally divided into regions of interests when monitoring cognitive tasks. The analysis was performed on a state-of-the-art workstation, allowing distributed computing to optimize the efficiency, and achieving a fit in less than one day for more than 20 million data points. This model consists of 1234 spatial parameters and 121824 temporal parameters (three for each location), thus achieving a compression rate of 3:100, which is vastly superior to traditional compression algorithms which can achieve

at most a 1:5 rate. The fit requires substantial computational power, but the estimates can then be used to generate surrogate runs almost instantaneously on a laptop.

Estimating all the parameters at once would require maximizing a likelihood over more than 100,000 parameters for more than 20 million data points, a extremely challenging task to perform within a reasonable time even with the most advanced computational facilities. Thus, we devised a step-wise estimation procedure with plug-in estimates from previous stages, which result in error propagation across stages that needs to be detected and mitigated. Bias propagation can be detected with diagnostic figures such as 3 and 4, and can be mitigated with an intermediate 2-band step before the latitudinal modeling to adjust the single band point estimates. Estimation uncertainty propagation in this context is of less concern, as given the considerable size of the data set, the estimated standard deviation is several orders of magnitude smaller than point estimates and the bias largely dominates the error propagation.

Despite the substantial improvements in flexibility, statistics-based compression is intrinsically dependent on the statistical model assumptions. The proposed methodology cannot generate surrogate runs that substitute the climate model, as the complex nonlinear dynamics of annual surface temperatures cannot be fully represented by a Gaussian process. As for emulators, our statistical model is to be regarded as a useful stochastic approximation that could help climate model users to test initial scientific hypotheses, but should not be used to perform a full geophysical investigation.

References

- Bhat, K., Haran, M., Olson, R., and Keller, K. (2012), “Inferring Likelihoods and Climate System Characteristics from Climate Models and Multiple Tracers,” *Environmetrics*, 23, 345–362.
- Bolin, D. and Lindgren, F. (2011), “Spatial models generated by nested stochastic partial

- differential equations, with an application to global ozone mapping,” *Annals of Applied Statistics*, 5, 523–550.
- Branstator, G. and Teng, H. (2010), “Two Limits of Initial-value Decadal Predictability in a CGCM,” *Journal of Climate*, 23, 6292–6311.
- Castruccio, S. and Genton, M. G. (2014), “Beyond Axial Symmetry: An Improved Class of Models for Global Data,” *Stat*, 3, 48–55.
- (2016), “Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature,” *Technometrics*, in press.
- Castruccio, S., McInerney, D. J., Stein, M. L., Liu, F., Jacob, R. J., and Moyer, E. J. (2014), “Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs,” *Journal of Climate*, 27, 1829–1844.
- Castruccio, S., Ombao, H., and Genton, M. G. (2016), “A Scalable Multi-Resolution Spatio-Temporal Model for Brain Activation and Connectivity in fMRI Data,” ArXiv: 1602.02435.
- Castruccio, S. and Stein, M. L. (2013), “Global Space-time Models for Climate Ensembles,” *Annals of Applied Statistics*, 7, 1593–1611.
- Chang, W., Haran, M., Olson, R., and Keller, K. (2015), “A Composite Likelihood Approach to Computer Model Calibration using High-dimensional Spatial Data,” *Statistica Sinica*, 25, 243–260.
- Collins, M. (2002), “Climate Predictability on Interannual to Decadal Time Scales: the Initial Value Problem,” *Climate Dynamics*, 19, 671–692.
- Collins, M. and Allen, M. R. (2002), “Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability,” *Journal of Climate*, 15, 3104–3109.
- Drignei, D., Forest, C. E., and Nychka, D. (2008), “Parameter Estimation for Computationally Intensive Nonlinear Regression with an Application to Climate Modeling,” *Annals of Applied Statistics*, 2, 1217–1230.

- Gent, P. R. et al. (2011), “The Community Climate System Model Version 4,” *J. Climate*, 24, 4973–4991.
- Genton, M. G., Castruccio, S., Crippa, P., Dutta, S., Huser, R., Sun, Y., and Vettori, S. (2015), “Visuanimation in statistics,” *Stat*, in press.
- Gneiting, T. (2013a), “Strictly and Non-strictly Positive Definite Functions on Spheres,” *Bernoulli*, 19, 1327–1349.
- (2013b), “Supplement to: Strictly and Non-strictly Positive Definite Functions on Spheres,” *Bernoulli*.
- Guinness, J. and Stein, M. (2013), “Transformation to Approximate Independence for Locally Stationary Gaussian Processes,” *Journal of Time Series Analysis*, 34, 574–590.
- Hansen, M. H. and Yu, B. Y. (2001), “Model Selection and the Principle of Minimum Description Length,” *Journal of the American Statistical Association*, 96, 746–774.
- Hitczenko, M. and Stein, M. (2012), “Some theory for anisotropic processes on the sphere,” *Statistical Methodology*, 9, 211 – 227, special Issue on Astrostatistics + Special Issue on Spatial Statistics.
- Holden, P. B. and Edwards, N. R. (2010), “Dimensionally Reduced Emulation of an AOGCM for Application to Integrated Assessment Modelling,” *Geophysical Research Letters*, 37.
- Holden, P. B., Edwards, N. R., Garthwaite, P. H., Fraedrich, K., Lunkeit, F., Kirk, E., Labriet, M., Kanudia, A., and Babonneau, F. (2013), “PLASIM-ENTSem: a Spatio-temporal Emulator of Future Climate Change for Impacts Assessment,” *Geoscientific Model Development Discussions*, 6, 3349–3380.
- Huang, C., Zhang, H., and Robeson, S. M. (2012), “A simplified representation of the covariance structure of axially symmetric processes on the sphere,” *Statistics and Probability Letters*, 82, 1346–1351.
- IPCC, . (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate*

- Change*, Cambridge University press, Cambridge, United Kingdom and New York, NY, USA: Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.).
- Jones, R. (1963), “Stochastic Processes on a Sphere,” *The Annals of Mathematical Statistics*, 34, 213–218.
- Jun, M. (2011), “Nonstationary Cross-covariance Models for Multivariate Processes on a Globe,” *Scandinavian Journal of Statistics*, 38, 726–747.
- (2014), “Matérn-based Nonstationary Cross-covariance Models for Global Processes,” *Journal of Multivariate Analysis*, 128, 134 – 146.
- Jun, M., Knutti, R., and Nychka, D. (2008), “Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There?” *Journal of the American Statistical Association*, 103, 934–947.
- Jun, M. and Stein, M. (2007), “An Approach to Producing Space x Time Covariance Functions on Spheres,” *Technometrics*, 49, 468–479.
- (2008), “Nonstationary Covariance Models for Global Data,” *Annals of Applied Statistics*, 2, 1271–1289.
- Lindgren, F., Rue, H., and Lindström, J. (2011), “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498.
- Lorenz, E. (1963), “Deterministic Nonperiodic Flow,” *Journal of the Atmospheric Sciences*, 20, 130–141.
- Poppick, A. and Stein, M. (2014), “Using Covariates to Model Dependence in Nonstationary, High-frequency Meteorological Processes,” *Environmetrics*, 25, 293–305.
- Priestley, M. B. (1965), “Evolutionary Spectra and Non-stationary Processes,” *Journal of the Royal Statistical Society. Series B*, 204–237.
- Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific.

- Sansó, B. and Forest, C. (2009), “Statistical Calibration of Climate System Properties,” *Journal of the Royal Statistical Society: Series C*, 58, 485–503.
- Sansó, B., Forest, C., and Zantedeschi, D. (2008), “Inferring Climate System Properties using a Computer Model,” *Bayesian Analysis*, 3, 1–37.
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *Annals of Statistics*, 6, 461–464.
- Taylor, K., Stouffer, R., and Meehl, G. (2012), “An Overview of CMIP5 and the Experiment Design,” *Bulletin of the American Meteorological Society*, 93, 485–498.
- Tukey, J. W. (1967), “An Introduction to the Calculations of Numerical Spectrum Analysis,” *Spectral Analysis of Time Series*, 25–46.
- Van Vuuren, D. et al. (2011), “The Representative Concentration Pathways: an Overview,” *Climatic Change*, 109, 5–31.